Low inter-rater reliability in traditional Chinese medicine for female infertility

Article *in* Acupuncture in Medicine · March 2011

DOI: 10.1136/aim.2010.003186 · Source: PubMed

3 authors:

Oddveig Birkeflet
University of Oslo
14 PUBLICATIONS **65** CITATIONS
SEE PROFILE

Petter Laake
University of Oslo
153 PUBLICATIONS **5,118** CITATIONS
SEE PROFILE

Nina Vøllestad
University of Oslo
131 PUBLICATIONS **6,396** CITATIONS
SEE PROFILE

Low inter-rater reliability in traditional Chinese medicine for female infertility

Oddveig Birkeflet, Petter Laake and Nina Vøllestad

Acupunct Med published online January 18, 2011
doi: 10.1136/aim.2010.003186

Low inter-rater reliability in traditional Chinese medicine for female infertility

Oddveig Birkeflet,[1] Petter Laake,[2] Nina Vøllestad[3]

[1]Institute of Health and Society, University of Oslo, Oslo, Norway
[2]Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway
[3]Institute of Health and Society, University of Oslo, Oslo, Norway

Correspondence to
Mrs Oddveig Birkeflet, Institute of Health and Society, University of Oslo, N-0318 Oslo, Norway;
oddveig.birkeflet@medisin.uio.no

Accepted 8 December 2010

ABSTRACT

Background Treatment of patients according to individual pattern diagnoses is an important feature of acupuncture rooted in traditional Chinese medicine (TCM). Little is known about the reliability of TCM pattern diagnoses.

Objective To examine in a cross-sectional study the inter-rater reliability of TCM diagnoses and acupuncture point selection.

Methods 30 infertile and 24 previously pregnant women were examined for TCM patterns by two acupuncturists. An operational interview guide related to gynaecology was used. The acupuncturists independently decided on the TCM patterns (categorised as excess, deficiency and merged patterns) and the prescription of acupuncture points. Kappa Statistics were used in the analyses.

Results 39 different TCM patterns and 36 different acupuncture points were used. For the choice of acupuncture points, poor to no agreement was found. Moderate to fair agreement was seen in excess/deficiency and merged patterns. Perfect match to moderate agreement on treatment was obtained when choosing meridians given certain TCM patterns.

Conclusions The low agreement on diagnoses indicates that acupuncturists follow individual pattern differentiation processes. The selection of acupuncture points seem to be closely related to the choice of TCM pattern diagnoses. The results indicate that the poor reliability in the diagnoses and thus treatment received by a patient will vary individually, which in turn is a challenge for clinical trials of acupuncture.

INTRODUCTION

Acupuncture is commonly used as an adjunct to in vitro fertilisation (IVF).[1] Although IVF and acupuncture is an active area of research,[1 2] the results of studies are difficult to interpret and compare because of a large variation in diagnostic criteria for patient inclusion and in the choice of acupuncture points.[1] This has been emphasised as a major obstacle for systematic reviews in this area.[3 4]

Individualised pattern diagnoses (based on signs and symptoms) and treatment according to their individual patterns are important features of acupuncture rooted in traditional Chinese medicine (TCM).[5–8] To ensure consistent and optimal treatment the pattern diagnoses must be reliable, but only a few studies have assessed the reliability of diagnostic data collected during a TCM examination.[9]

O'Brien and Birch reviewed studies of the reliability of traditional East Asian medicine diagnoses and concluded that reliability of pattern diagnoses and treatment was low.[10] Reliability has been studied for different patient groups using a wide range of research designs. Some studies have examined how several clinicians diagnose and suggest treatment for one or only a few patients,[11 12] whereas others have compared pairs of clinicians examining a number of patients.[13] The results of these studies show no consistent pattern. A relatively good agreement has been reported for TCM diagnoses while suggestions for acupuncture points have shown greater variation.[12] Most of the studies report percentage agreement and sometimes the correlation between raters. Relevant statistical analyses such as Kappa (κ) statistics are rarely reported.

Decisions about treatments should be based on the TCM pattern diagnoses and hence it is possible that low agreement between clinicians for suggested acupuncture points might be caused by differences in diagnoses. Agreement on the relationship between diagnoses and acupuncture treatment has to our knowledge never been investigated. In one study on low back pain, Hogeboom *et al*[11] reported insignificant correlations between diagnoses and acupuncture point selection by six acupuncturists. κ Statistics were not reported.

Several factors may potentially affect the poor reliability in TCM pattern diagnoses and suggestions for treatment. The practitioners may vary owing to differences in clinical education and experience. Furthermore, studies have often used a design in which patients are examined in sequence by the acupuncturists. If more than one acupuncturist interviews the same patient at different times, this may induce variations in the way in which the patient presents their symptoms and signs. A low agreement may thus not reflect differences between clinicians, but variations that can be attributed to the patients. There is a need for studies that reduce this effect, in order to examine the true inter-rater differences.

The studies so far have examined the inter-rater reliability of TCM patterns in specific patient groups where the variation in signs and symptoms might be small. Hence, the expectations of the practitioner may influence the examination and the conclusions about TCM patterns. The variability may then be underestimated and reliability higher than in a normal clinical setting. In our study, we wanted to increase the variability and thus strengthen the test by examining the reliability in a mixed group of fertile and infertile woman. Furthermore, different biomedical causes of infertility were included for the same reason.

The objective of this study was to determine the inter-rater reliability in TCM patterns and prescriptions of acupuncture points.

METHODS

Study design

This study was designed as a cross-sectional study of two groups of women: infertile and fertile. For this analysis, the participants were combined into one group. For two acupuncturists we examined three aspects of the inter-rater reliability: diagnoses of TCM patterns, single acupuncture point selection and point prescription according to TCM patterns.

Participants

Participants were 54 Norwegian-speaking women with an average age of 33.3 years (range 24–42). In the period from September 2007 to October 2008, 30 infertile and 24 women who had previously achieved spontaneous pregnancy were consecutively interviewed. The infertile women were recruited among those included in an IVF programme and they met the medical requirement for infertility diagnosis—a failure to conceive after 12 months of unprotected intercourse.[14] Twenty-four of them were primary infertile (never been pregnant) and six had secondary infertility (had children in the past). Eleven were still under medical examination and were regarded as unexplained infertility. The self-reported biomedical diagnoses were endometriosis (n=5), polycystic ovarian syndrome (n=6) and poor egg quality (n=3). Fertile women were recruited from women who had previously been spontaneously pregnant. They had delivered within the last 3–12 months before participating. Participants were recruited via advertisements on websites for maternity care and posters displayed at doctors' offices. All participants volunteered for the study and signed a written informed consent.

Setting

The interview took place in an acupuncture clinic in Asker, Norway. For each participant the two acupuncturists attended the consultation together, to ensure simultaneous access to the information. This minimised the potential for observational changes/bias. The consultation was based on the four diagnostic methods of inquiry as described by Maciocia: case history taking, palpation, observation and auscultation.[15] An operational structured interview guide according to Maciocia's symptoms and signs in gynaecology[15] ensured that all the participants were asked identical questions. Supplementary information was collected according to individual symptoms. While one acupuncturist guided the interview, the other listened and had an opportunity to make notes and ask additional questions. Both acupuncturists concurrently examined the tongue and the radial pulse bilaterally on each participant. They did not discuss their findings with each other. After completing the interview, the acupuncturists individually diagnosed the TCM patterns based on their own judgement and criteria. Finally, they provided the patterns with recommended acupuncture points. The acupuncturists were aware of whether the woman examined was fertile or infertile. The use of a structured interview guide resulted in interviews lasting about 60 min.

Both acupuncturists were educated at a Norwegian acupuncture college offering a bachelor's degree in TCM acupuncture. One acupuncturist had 6 years and the other 20 years of clinical experience plus an advanced course from Nanjing College of TCM in Nanjing China. One of the acupuncturists participated in the research team.

Data analysis

As several TCM patterns were used only once, we analysed the patterns in merged groups. The single TCM patterns from each acupuncturist were merged into excess and deficiency patterns as illustrated in figure 1. To examine if the agreement improved when merged further on a higher level, we united the excess and deficiency patterns from the same organ system to the merged patterns. The pattern categories consist of dichotomous variables (figure 1).

One pattern identifies imbalances in two organs: Heart and Spleen blood deficiency. This pattern was categorised under Heart deficiency.

To examine if the agreement for recommended treatment improved on a higher level, the single acupuncture points on the same meridian were merged. The frequency of agreement on diagnosis or acupuncture points (or their merged groups) is termed 'mutual positive score'.

Statistical methods

To examine inter-rater agreement the κ statistic was used to assess the level of agreement between two acupuncturists beyond that expected by chance. κ values <0.20 were considered as poor agreement, 0.21–0.40 as fair, 0.41–0.60 as moderate and 0.61–0.80 were considered as good. Values between 0.81 and 1.00 were regarded as very good agreement.[16] The marginal totals for the 2×2 table are not balanced, the observed proportion of agreement is quite high, but the value of κ indicates a low level of reliability. This is a known paradox of the κ statistic. The κ statistic alone is insufficient. Therefore, we also report the maximum κ value. The maximum

agreement for κ is 1.00, a perfect agreement and 0 indicates agreement no better than a chance. Negative values show a worse than chance agreement.[16] The calculations were done in SPSS 16.0 for Windows. The 95% CI and maximum κ were calculated with DAG_Stat.[17]

RESULTS

Altogether 39 different TCM patterns were used; acupuncturist 1 (acu1) used 32 and acupuncturist 2 (acu2) used 29 different patterns. Most often, several patterns were diagnosed on each participant, on average acu1 set six and acu2 set five patterns. A total of 36 different acupuncture points were used. Acu1 used 34 and acu2 used 22 different points. On average, they selected eight and six points, respectively, for individual subjects.

The reliability of TCM patterns and acupuncture points was analysed separately and thereafter the reliability of

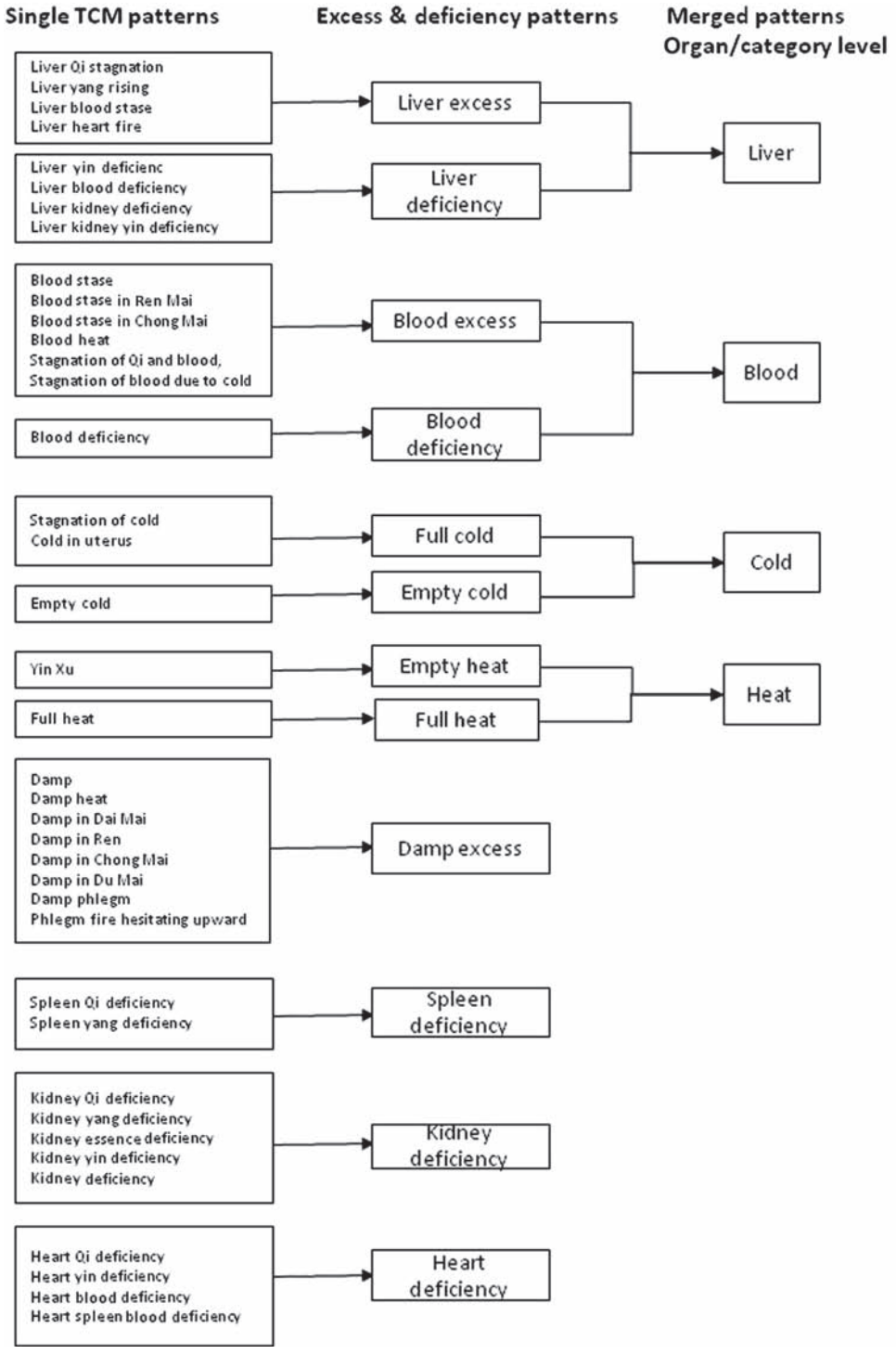

Figure 1 Definition of the variables when merging single patterns into excess and deficiency patterns and further to organ/category level.

acupuncture points for given merged TCM patterns was analysed.

TCM patterns

Damp excess pattern was diagnosed for 21 and 27 women by the two acupuncturists, respectively. There was a fair agreement for this category (table 1). The Liver excess pattern was diagnosed for most of the participants and the maximum κ indicates a moderate agreement (table 1). Spleen and Kidney deficiency were the most commonly used deficiency patterns. Maximum κ showed a fair agreement for both patterns. Among the merged patterns, Liver was the most used pattern, in 94% of the respondents. It was the pattern with the highest inter-rater agreement on diagnoses; maximum κ indicated a moderate agreement. The cases for which κ showed no agreement and κ maximum showed a fair agreement were interpreted as agreement of not using these patterns, since the patterns were used infrequently.

The acupuncture points

For the least used points, the maximum κ values indicate a fair to moderate agreement (table 2). Since these points were seldom in use, this probably reflects agreement of not using these points.

The most commonly used point was LR3, which was used on almost all the women. The negative κ value is due to imbalance in the distribution of marginal totals in the 2×2 matrix used to calculate the κ value. However, maximum κ shows a fair agreement. For KI3 and SP6, which were used on a majority of the women, the agreement was much lower. For the other points the two acupuncturists differed to a large extent in the frequency of their use, with poor to fair agreement based on maximum κ.

The meridians

Agreement was also examined after merging the points for each meridian—for example, all the Liver points were collected and named as the Liver meridian (table 3). In general, the maximum κ values were about the same as for single points.

For the most often used meridians (Liver, Kidney, Stomach and Spleen) the acupuncturists showed a moderate to fair agreement (maximum κ). Also for the least used meridians, a fair agreement was seen (maximum κ). This again may reflect agreement in not choosing the points on the meridian. For the other meridians the agreement was poor.

Meridians for given merged patterns

In those cases where the two acupuncturists agreed on the merged patterns, almost complete agreement was seen for choosing the Liver meridian (table 4). Furthermore, the Liver meridian was selected for almost all women and consequently, the Liver meridian was found in combination with nearly all other patterns. Hence, a high agreement was seen for Liver meridian at all patterns. For the other meridians fair agreement was seen in the use on the merged patterns. κ could not be calculated in some cases because acu2 had used the meridian in all cases.

DISCUSSION

Our findings showed moderate agreement for the Liver patterns and fair to poor agreement for the other patterns. For the most used acupuncture points the agreement was fair and fair to moderate agreement for the meridian level. For the selection of meridians to use in the treatment of given merged TCM patterns, we found 100% agreement in using the Liver meridian on 50% of the patterns and the

Table 1 Frequencies of the merged traditional Chinese medicine patterns diagnosed by each acupuncturist, their mutual positive score, κ, CI and the maximum κ value

Patterns (n = 54)	Acu1 (n)	Acu2 (n)	Mutual positive score	κ	κ (95% CI)	Max κ
Excess patterns						
Damp excess	21	27	14	0.22	−0.03 to 0.48	0.25
Liver excess	51	47	45	0.13	−0.21 to 0.47	0.42
Blood excess	41	9	8	0.06	−0.04 to 0.17	0.09
Full heat	3	9	0	−0.09	−0.17 to −0.01	0.37
Full cold	6	6	0	−0.13	−0.20 to −0.01	0.37
Deficiency patterns						
Heart deficiency	6	3	2	0.4	−0.02 to 0.82	0.45
Blood deficiency	3	5	1	0.19	−0.22 to 0.61	0.44
Spleen deficiency	46	42	37	0.15	−0.15 to 0.44	0.34
Liver deficiency	11	20	5	0.08	−0.17 to 0.33	0.25
Kidney deficiency	44	37	30	−0.01	−0.26 to 0.23	0.25
Empty heat	12	3	0	−0.10	−0.19 to -0.00	0.33
Empty cold	0	4	0			
Merged patterns						
Liver	51	51	49	0.30	−0.21 to 0.80	0.46
Heat	15	12	5	0.16	−0.12 to 0.45	0.20
Blood	42	12	12	0.15	0.04 to 0.26	0.13
Cold	6	10	2	0.13	−0.18 to 0.43	0.40

Birkeflet O, Laake P, Vøllestad N. *Acupunct Med* (2011). doi:10.1136/aim.2010.003186

Table 2 Distribution of the acupuncture points and reliability measures. The frequency of each acupuncturist's use of points, the mutual positive score, κ, CI and the maximum κ value

Point	Acu1 (n)	Acu2 (n)	Mutual positive score	κ	κ (95% CI)	Max κ
The less used points						
LR13	1	8	1	0.20	−0.13 to 0.53	0.43
KI13	1	6	0	−0.03	−0.09 to 0.02	0.43
PC6	9	2	0	−0.06	−0.14 to 0.01	0.38
KI7	8	1	1	0.20	−0.13 to 0.53	0.43
GB34	3	8	1	0.11	−0.20 to 0.42	0.40
The most used points						
LR3	51	47	44	−0.08	−0.16 to −0.01	0.39
KI3	38	38	25	−0.16	−0.40 to 0.09	0.18
SP6	35	45	29	−0.02	−0.25 to 0.22	0.23
Point used differently						
ST40	31	5	1	−0.12	−0.28 to 0.06	0.09
SP3	9	28	5	0.02	−0.17 to 0.22	0.17
ST36	17	34	12	0.09	−0.13 to 0.30	0.17
SP9	4	20	4	0.24	0.04 to 0.44	0.31
LR8	9	23	3	−0.07	−0.28 to 0.14	0.18
CV4	27	13	8	0.11	−0.12 to 0.34	0.21
SP10	12	8	3	0.15	−0.15 to 0.44	0.34
LI11	7	17	3	0.08	−0.16 to 0.33	0.29

Table 3 Frequencies of the acupuncturists' use of the most used meridians, their mutual positive score, κ, CI and the maximum κ value

Meridian	Acu1 (n)	Acu2 (n)	Mutual positive score	κ	κ (95% CI)	Max κ
Liver	52	52	50	−0.04	−0.00 to −0.01	0.46
Kidney	51	40	38	0.03	0.23 to −0.17	0.33
Stomach	50	39	35	−0.13	−0.02 to −0.24	0.27
Spleen	45	51	42	−0.09	−0.01 to −0.17	0.37
Large intestine	25	27	8	0.01	0.26 to −0.24	0.18
Conception vessel	27	16	8	0.00	0.24 to −0.24	0.17
Heart	14	1	1	0.10	0.29 to −0.10	0.35
Gallbladder	7	8	1	−0.01	0.26 to −0.27	0.35
Lung	9	1	0	−0.04	0.03 to −0.10	0.39

data show moderate to fair agreement for the other patterns.

The chosen design with simultaneous participation of the consultation did not allow more than two acupuncturists. However, this ensured that the two acupuncturists simultaneously accessed the same clinical information. Hence, the differences in diagnostics must arise in the interpretation process from symptoms and signs to conclusion about the TCM pattern diagnoses.

Previous studies have used a different design, where the patients are examined repeatedly.[11 12 18–21] Hence, one cannot distinguish between differences owing to presentations of symptoms and differences that can be ascribed to the clinician's interpretation.

One possible explanation for different interpretation may be differences in their background and clinical practice. However, the acupuncturists had their education from the same school and foolowed the same curriculum. Their clinical practice was similar, although one of them had greater experience as acupuncturist (20 vs 6 years). Previous studies show that acupuncturists possessing at least a bachelor's degree and a minimum of 5 and up to 20 years of experience, report poor consistency in agreement on TCM pattern diagnoses.[9 11 13 18–24] Hence, it seems that low agreement is typical even among those with a long education and clinical experience and not caused by the design of this study.

This study showed that even when TCM patterns were collapsed into broader categories, such as the merged patterns, higher inter-rater reliability was not achieved. This finding is consistent with the results of Mist *et al*, who found no improvement in agreement when they united TCM patterns into broader categories.[13] Our study partly followed the same procedure in grouping patterns.

One complication in the process of differentiating symptoms and signs into TCM patterns is that some symptoms and signs observed together may provide conflicting information. In gynaecology, Maciocia states that conflicting and contradictory gynaecological manifestations occur commonly.[15] Yet TCM textbooks lack clear guidelines for interpreting contradictory information.[6 15 25] One possible

Table 4 Frequencies of the meridians of the Liver, Spleen, Kidney and Stomach used on the patterns of the Liver, Spleen, Kidney, Damp and Blood, the mutual positive score and the κ, CI and the maximum κ value

Meridian	Acu1 (n)	Acu2 (n)	Mutual positive score	κ	κ (95% CI)	Max κ
Liver merged pattern (n=49)						
Liver meridian	47	49	47	†		
Spleen meridian	40	46	37	−0.10	−0.19 to −0.01	0.35
Kidney meridian	47	37	36	0.08	−0.15 to 0.31	0.35
Stomach meridian	45	34	30	−0.15	−0.27 to −0.02	0.25
Liver excess (n=45)						
Liver meridian	43	45	43	†		
Spleen meridian	37	42	34	−0.11	−0.20 to −0.01	0.35
Kidney meridian	43	34	33	0.09	−0.16 to 0.33	0.35
Stomach meridian	41	32	28	−0.16	−0.29 to −0.03	0.25
Liver deficiency (n=5)						
Liver meridian	5	5	5	*		
Spleen meridian	4	4	3	−0.25	−0.59 to 0.09	0.24
Kidney meridian	4	3	3	0.55	−0.16 to 1.26	
Stomach meridian	4	4	3	−0.25	−0.59 to 0.09	0.24
Spleen deficiency (n=37)						
Liver meridian	37	35	35	†		
Spleen meridian	32	37	32	†		
Kidney meridian	35	26	25	0.07	−0.17 to 0.31	0.31
Stomach meridian	34	26	23	−0.15	−0.29 to −0.01	0.25
Kidney deficiency (n=30)						
Liver meridian	29	29	28	−0.03	−0.08 to 0.01	0.47
Spleen meridian	23	28	21	−0.12	−0.25 to 0.02	0.31
Kidney meridian	29	30	29	–		
Stomach meridian	27	23	20	−0.16	−0.30 to −0.02	0.29
Damp excess (n=14)						
Liver meridian	12	14	12	†		
Spleen meridian	11	14	11	†		
Kidney meridian	14	10	10	†		
Stomach meridian	13	10	9	−0.13	−0.35 to 0.09	0,27
Blood (n=12)						
Liver meridian	12	12	12	*		
Spleen meridian	10	12	10	†		
Kidney meridian	11	9	8	−0.14	−0.37 to 0.08	0.29
Stomach meridian	7	7	7	†		
Blood excess (n=8)						
Liver meridian	8	8	8	*		
Spleen meridian	6	8	6	†		
Kidney meridian	7	7	6	−0.14	−0.34 to 0.05	0.35
Stomach meridian	8	5	5	†		

The blank fields show that the κ could not be calculated owing to the frequency distribution in the 2 × 2 matrices.

*100% agreement; †one of the acupuncturists has used the meridian in all the cases diagnosed with the pattern.

cause for the differences in diagnoses may thus be the different importance and weight put on the different observations. Mist *et al* found that the inter-rater reliability of TCM diagnoses was improved through a calibration training process among practitioners when they used a questionnaire-based diagnoses process designed to cover the major factors for pattern differentiations.[13] Their results indicate that clear guidelines and definitions lead to greater consensus among the practitioners. This study only used a common interview guide to obtain the same procedure for obtaining data information. No attempts were used to standardise the interpretation process for signs and symptoms of TCM diagnosis.

The reliability was also poor for the acupuncture points, even when we merged them according to the meridians. Similarly, Hogeboom *et al* found very low agreement among practitioners about which patients should receive a given acupuncture point.[11] However, they also examined the acupuncture points for relationship with specific TCM diagnoses. By grouping some of the points into 12 clusters, they found that only two clusters were strongly associated with a specific diagnosis.[11] We found improved reliability when the acupuncture points were categorised to meridians and examined for a given merged pattern diagnosis. Thus, it seems that the two acupuncturists achieved greater consensus about the treatment for a

given diagnosis than in making the diagnoses. Correspondingly, Zhang *et al*, found that practitioners had better agreement about a diagnosis and the herbal prescription required, than agreement between the practitioners' prescriptions.[21] This is in keeping with the textbooks and education that provide recommendations about which treatment/acupuncture points to use on specific patterns.[25]

CONCLUSION

We found unsatisfactory low inter-rater reliability in the individualised TCM pattern diagnoses and also in the selection of acupuncture points. The low agreement about diagnoses and acupuncture point selection indicate that acupuncturists follow individual pattern differentiation processes. This leads to differences in treatment of similar conditions, which in turn is a challenge for clinical trials of acupuncture. The key seem to be in the diagnostic process, since there is a strong relationship between TCM pattern diagnosis and selection of acupuncture points.

Summary points

- Previous studies have mostly found that TCM diagnosis has poor reliability.
- We compared diagnosis and point selection made by two acupuncturists at the same consultation.
- Agreement was generally unsatisfactory, though moderate for selecting points for a given diagnosis.

Acknowledgements The authors acknowledge Christina Weseth for providing study data and they also thank the participating women.

Funding Norwegian Acupuncturist Assosiation, The National Research Center in Complementary and Alternative Medicine through The Norway–China Cooperation funds and Pharma West AS.

Competing interests None.

Patient consent Obtained.

Ethics approval This study was conducted with the approval of the Regional Committee for Medical and Health Research Ethics.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. Stener-Victorin E, Humaidan P. Use of acupuncture in female infertility and a summary of recent acupuncture studies related to embryo transfer. *Acupunct Med* 2006;24:157–63.
2. White AR. A review of controlled trials of acupuncture for women's reproductive health care. *J Fam Plann Reprod Health Care* 2003;29:233–6.
3. Ng EH, So WS, Gao J, *et al.* The role of acupuncture in the management of subfertility. *Fertil Steril* 2008;90:1–13.
4. NIN Consensus Development Panel on Acupuncture. Acupuncture. *JAMA* 1998;280:1518–24.
5. Zhen H, Fei-xia D. *Clinical Reasoning in Chinese Medicine*. China: People's Medical Publishing House 2008.
6. Kaptchuk T. *Chinese Medicine: The Web That Has No Weaver*. London: Rider 1987.
7. Liu T. Role of acupuncturists in acupuncture treatment. *Evid Based Complement Alternat Med* 2007;4:3–6.
8. El-Toukhy T, Sunkara SK, Khairy M, *et al.* A systematic review and meta-analysis of acupuncture in *in vitro* fertilisation. *BJOG* 2008;115:1203–13.
9. Sherman KJ, Cherkin DC, Hogeboom CJ. The diagnosis and treatment of patients with chronic low-back pain by traditional Chinese medical acupuncturists. *J Altern Complement Med* 2001;7:641–50.
10. O'Brien KA, Birch S. A review of the reliability of traditional East Asian medicine diagnoses. *J Altern Complement Med* 2009;15:353–66.
11. Hogeboom CJ, Sherman KJ, Cherkin DC. Variation in diagnosis and treatment of chronic low back pain by traditional Chinese medicine acupuncturists. *Complement Ther Med* 2001;9:154–66.
12. Kalauokalani D, Sherman KJ, Cherkin DC. Acupuncture for chronic low back pain: diagnosis and treatment patterns among acupuncturists evaluating the same patient. *South Med J* 2001;94:486–92.
13. Mist S, Ritenbaugh C, Aickin M. Effects of questionnaire-based diagnosis and training on inter-rater reliability among practitioners of traditional Chinese medicine. *J Altern Complement Med* 2009;15:703–9.
14. World Health Organization. Assisted Reproductive Technologies (ARTs), 2010. http://www.who.int/genomics/gender/en/index6.html (accessed 1 Jun 2010).
15. Maciocia G. *Obstetrics & Gynecology in Chinese Medicine*. London: Churchill Livingstone 1998.
16. Altman DG. *Practical Statistics for Medical Research*. London: Chapman & Hall/CRC 1999.
17. Mackinnon A. DAG_Stat, 2000. http://www.mhri.edu.au/biostats/DAG_Stat/ (accessed 1 Jan 2010).
18. Coeytaux RR, Chen W, Lindemuth CE, *et al.* Variability in the diagnosis and point selection for persons with frequent headache by traditional Chinese medicine acupuncturists. *J Altern Complement Med* 2006;12:863–72.
19. O'Brien KA, Abbas E, Zhang J, *et al.* An investigation into the reliability of Chinese medicine diagnosis according to Eight Guiding Principles and Zang-Fu Theory in Australians with hypercholesterolemia. *J Altern Complement Med* 2009;15:259–66.
20. Sung JJ, Leung WK, Ching JY, *et al.* Agreements among traditional Chinese medicine practitioners in the diagnosis and treatment of irritable bowel syndrome. *Aliment Pharmacol Ther* 2004;20:1205–10.
21. Zhang GG, Lee W, Bausell B, *et al.* Variability in the traditional Chinese medicine (TCM) diagnoses and herbal prescriptions provided by three TCM practitioners for 40 patients with rheumatoid arthritis. *J Altern Complement Med* 2005;11:415–21.
22. Kim M, Cobbin D, Zaslawski C. Traditional Chinese medicine tongue inspection: an examination of the inter- and intrapractitioner reliability for specific tongue characteristics. *J Altern Complement Med* 2008;14:527–36.
23. O'Brien KA, Abbas E, Zhang J, *et al.* Understanding the reliability of diagnostic variables in a Chinese Medicine examination. *J Altern Complement Med* 2009;15:727–34.
24. Sherman KJ, Hogeboom CJ, Cherkin DC. How traditional Chinese medicine acupuncturists would diagnose and treat chronic low back pain: results of a survey of licensed acupuncturists in Washington State. *Complement Ther Med* 2001;9:146–53.
25. Maciocia G. *The Practice of Chinese Medicine: The Treatment of Diseases with Acupuncture and Chinese Herbs*. London: Churchill Livingstone 1994.